Publish to CDH-HDFS How to Configure

On this page

Step 1. Launch Aspire and Open the Content Source Management Page

? Unknown Attachment

Launch Aspire (if it's not already running). See:

- Launch Control
- Browse to: http://localhost:50505. For details on using the Aspire Content Source Management page, please refer to Admin UI

Step 2. Add a New Content Source

For this step please follow the step from the Configuration Tutorial of the connector of you choice, please refer to Connector list

Step 3. Add a New Publish to CDH-HDFS to the Workflow

To add a Publish to CDH-HDFS drag from the **Publish to CDH-HDFS** rule from the *Workflow Library* and drop to the *Workflow Tree* where you want to add it. This will automatically open the Publish to CDH-HDFS window for the configuration of the publisher.

1. Enter the name of the publisher. *(This name must be unique)*.
2. Enter the description of the publisher that will be shown in the Workflow Tree.
3. Select the publishing protocol to use:
   a. HDFS (Java API)
   b. WebHDFS (REST API)

⚠ Not all HDFS clusters have WebHDFS enabled.

3.a. Publish Using HDFS

In the **HDFS** section of the *Publish to CDH HDFS* window specify the connection information to publish to HDFS.

1. Enter the *HDFS URL*. Use **hdfs://** protocol and the port (by default 8020). I.e. hdfs://localhost:8020
2. Specify the location of the *Output key*. An AXPath of the node inside the AspireObject. I.e. /doc/docType
3. Specify the absolute *HDFS Folder Path* where the files will be published to. I.e. /user/jsmith/my_aspire_output. (The user which runs Aspire must have write access to the HDFS folder).
4. Specify the *Max File Size* in MegaBytes. If left as -1 it will use the HDFS Block Size as the file limit.
5. Specify a *File Prefix Name*. I.e. *aspire-*, files will be named: aspire-00000, aspire-00001, aspire-00002m, etc.
6. Debug: Check if you want to run the publisher in debug mode.
7. Click ***Add***.

Once you've clicked on the *Add* button, it will take a moment for Aspire to download all of the necessary components (the Jar files) from the Maven repository and load them into Aspire. Once that's done, the publisher will appear in the Workflow Tree.

ⓘ For details on using the Workflow section, please refer to the Workflow introduction.